

City University of New York (CUNY)

CUNY Academic Works

Open Educational Resources

City College of New York

2018

Lecture: Intro to Data Science - ML 4 - Week Twelve

Grant Long
CUNY City College

NYC Tech-in-Residence Corps

[How does access to this work benefit you? Let us know!](#)

More information about this work at: https://academicworks.cuny.edu/cc_oers/186

Discover additional works at: <https://academicworks.cuny.edu>

This work is made publicly available by the City University of New York (CUNY).
Contact: AcademicWorks@cuny.edu

City College, Fall 2018

Intro to Data Science

Week 12: Unsupervised Learning

November 26, 2018

Today's Agenda

1. Supervised vs. Unsupervised Learning
2. Clustering
3. Topic Modeling
4. Hacking Visualizations in Python

Supervised vs. Unsupervised Learning

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.

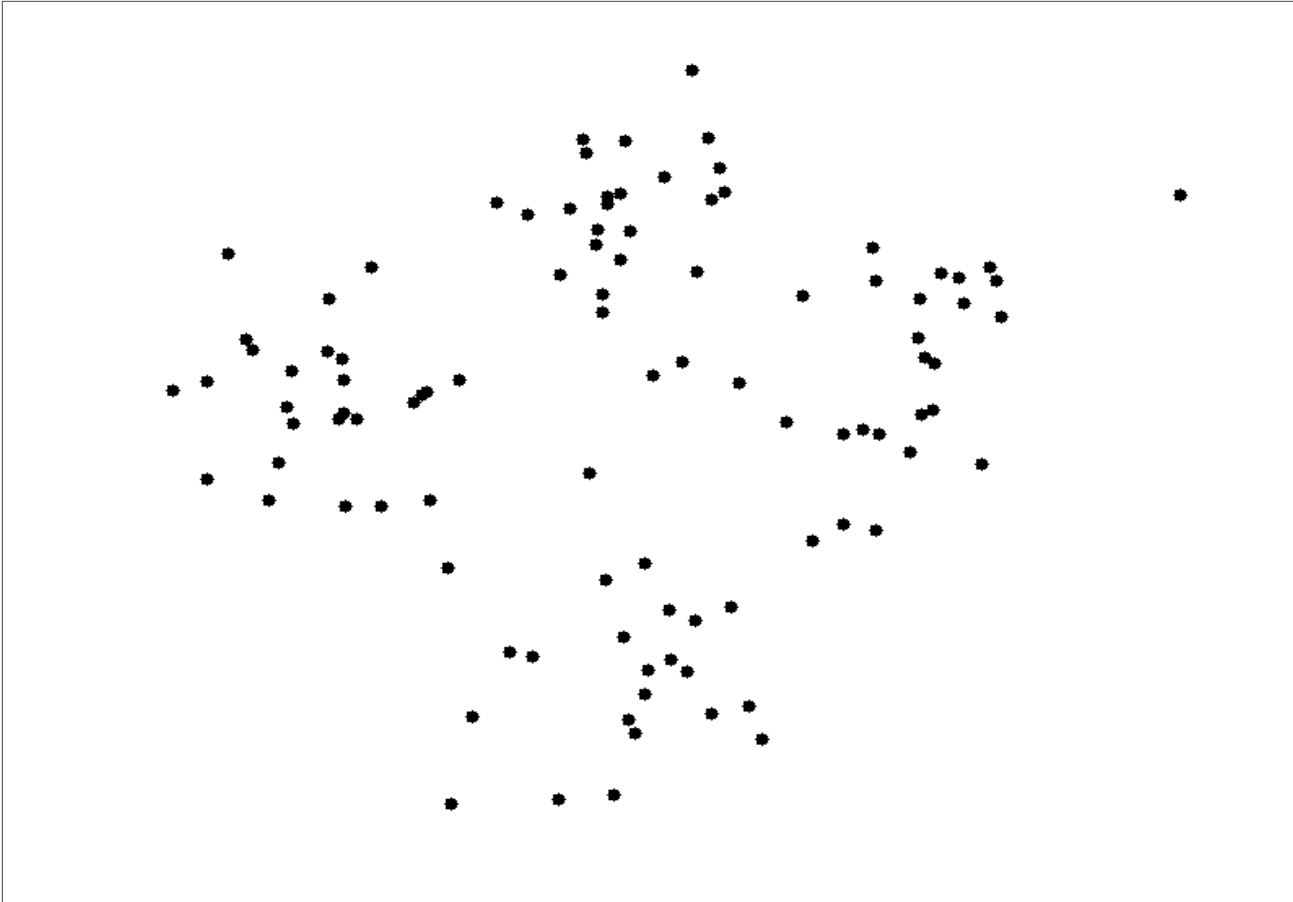
Unsupervised learning is a branch of machine learning that learns from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data.

Clustering

Common Clustering Techniques

- K-means
- Mean shift
- Hierarchical

How would you group these points?



Visualization by Andrey A. Shabalin.

K-Means Clustering

1. Choose k , a number of clusters.
2. Pick k starting points.
3. Assign each point to a cluster based on the closest of the k chosen points.
4. Calculate the points at the center of each cluster.
5. Assign each point to a cluster based on the closest of the new center points.
6. Repeat steps 4 & 5 until convergence.

For more, check out these **these help notes** from CS221 at Stanford.

K-Means: First Try



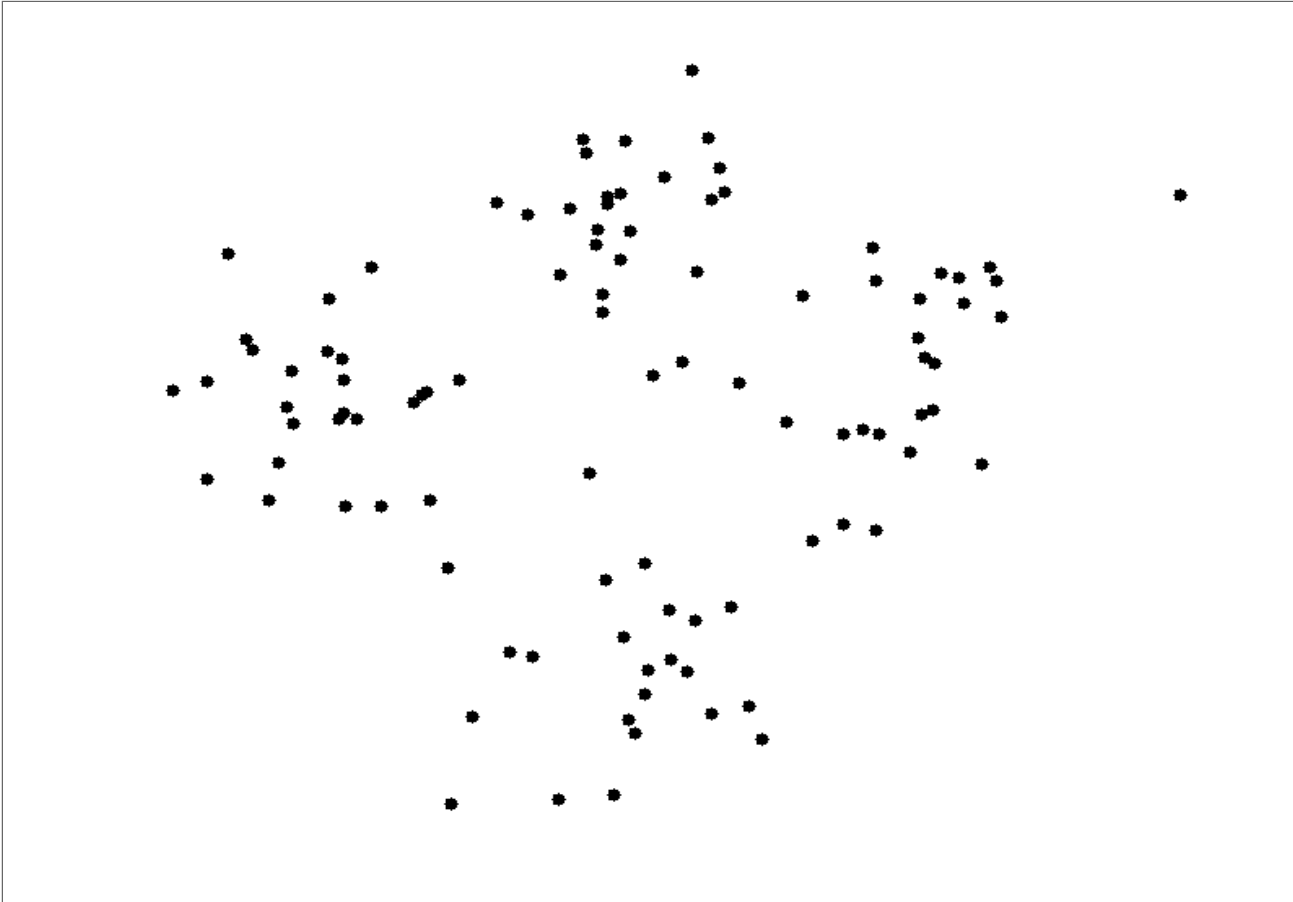
Visualization by Andrey A. Shabalin.

K-Means: Second Try



Visualization by Andrey A. Shabalin.

K-Means: Third Try



Visualization by Andrey A. Shabalin.

K-Means: Fourth Try



Visualization by Andrey A. Shabalin.

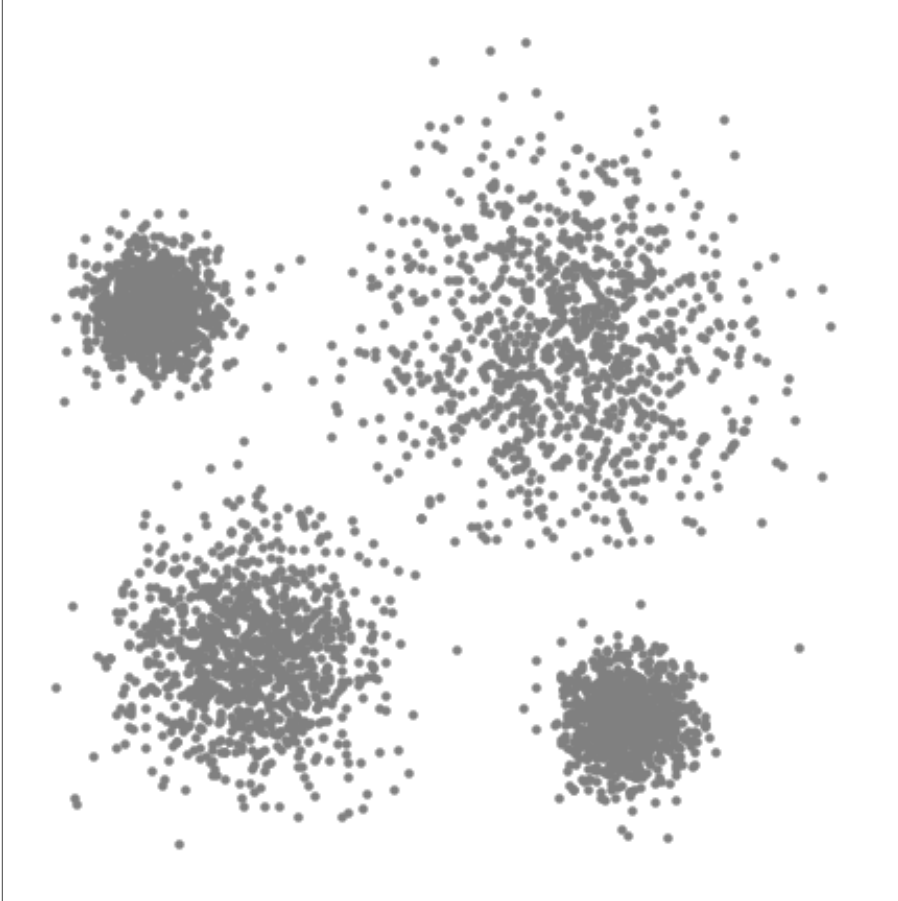
Demo

Mean Shift Clustering

1. Arrange windows to cover all points.
2. Compute the number of points in each frame.
3. Shift the window to the mean.
4. Repeat until convergence.

These slides adapted from **CS109 at Harvard**.

Mean Shift Clustering



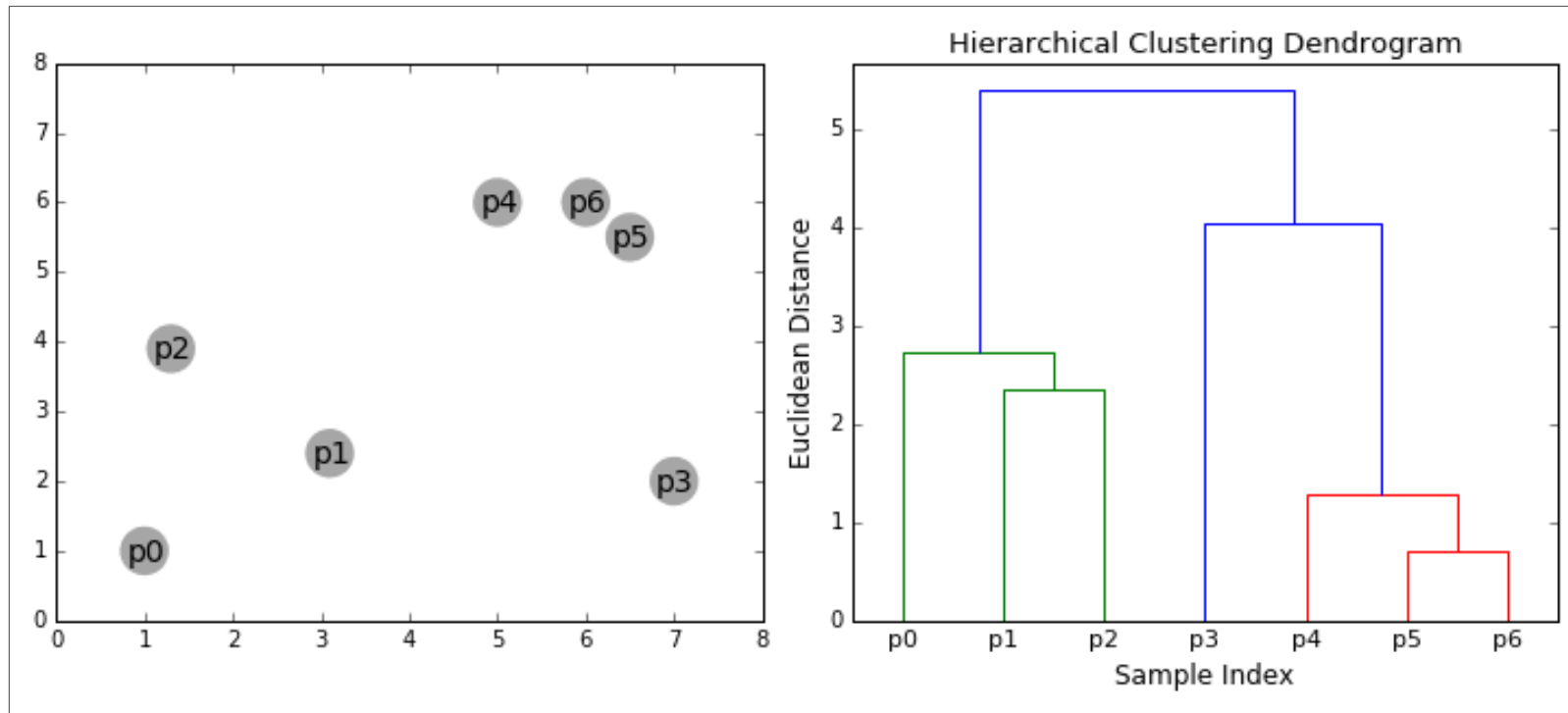
Visualization by David Sheehan.

Hierarchical Clustering

1. Each observation starts in its own cluster.
2. The two closest pairs form a cluster.
3. Clusters are merged as one moves up the hierarchy.
4. Repeat until all points belong to the same cluster.

See **Wikipedia** for more.

Hierarchical Clustering



Visualization by David Sheehan.

Clustering Applications

- Google Image Search Categories
- Author Clustering
- Picking Locations for Hospitals, Police Stations, etc.
- Outlier Detection

These slides adapted from **CS109 at Harvard**.

Clustering Application Example

Multi-feature trajectory clustering using mean shift



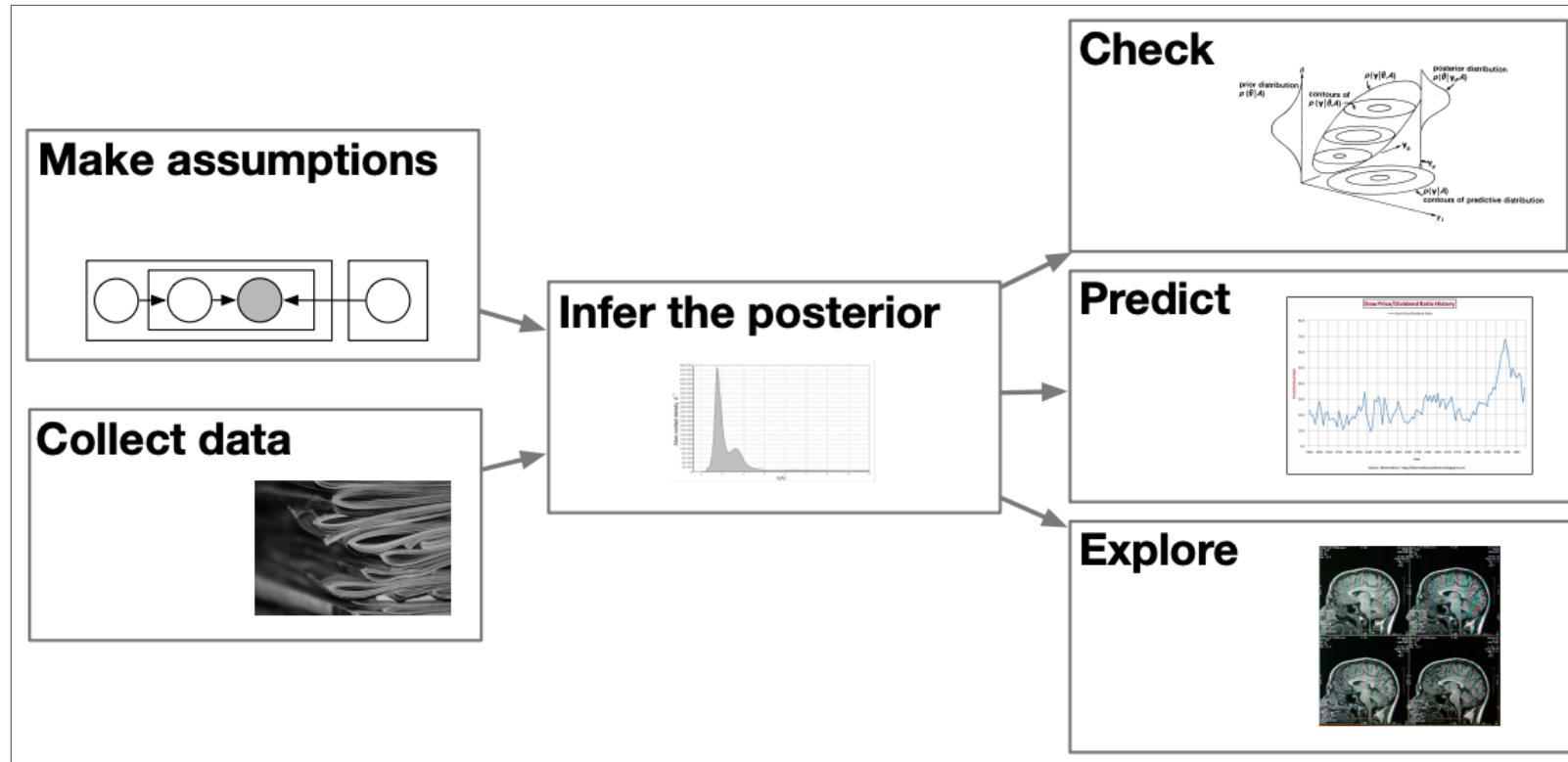
Topic Modeling

Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

- Discover the hidden themes that pervade the collection.
- Annotate the documents according to those themes.
- Use annotations to organize, summarize, and search the texts.

This slide adapted from **Columbia's David Blei**.

Latent Dirichlet Process for Topic Modeling



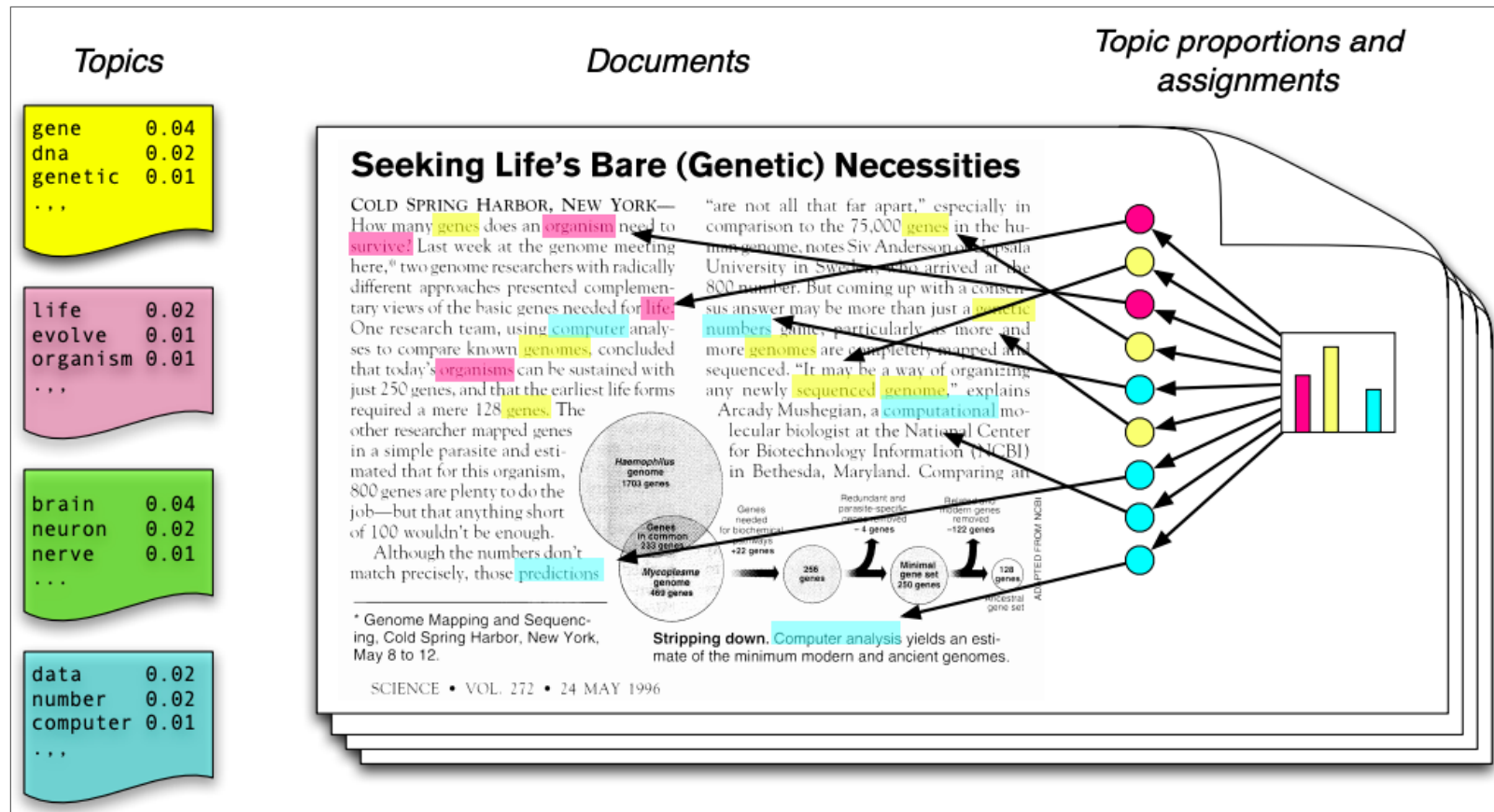
This slide adapted from Columbia's David Blei.

What topics does this passage cover?

How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job - but that anything short of 100 wouldn't be enough.

This slide adapted from **Columbia's David Blei**.

Assigning Topics Via Machine Learning

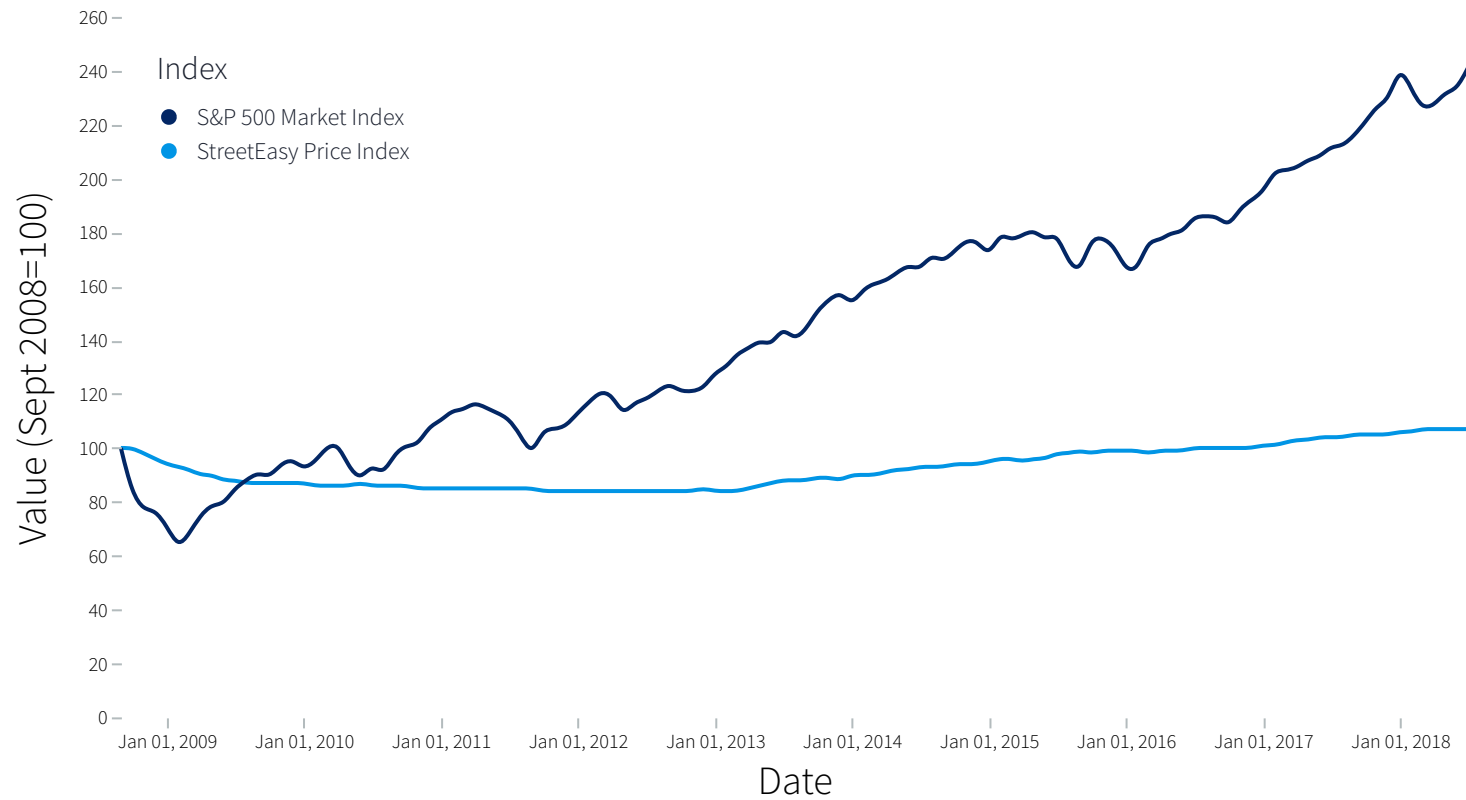


This slide adapted from Columbia's David Blei.

Hacking Visualizations in Python

Using **Altair** to Create Interactive d3.js Visualizations

StreetEasy Price Index vs. S&P 500



Topic Modeling is Hard!

We'll use **Gensim** to build our topic model and **pyLDAvis** to visualize it.

Reminder: Assignment 7: Due Monday, December 3 by 6:30pm

DataCamp's Deep Learning in Python

- The course should appear collectively as assignment within your existing DataCamp account.
- Each section will appear separately and will be worth one point toward the total grade for the homework, plus an additional point for overall effort.
- Course claims to take 4 hours - as always, use your time wisely.
- This will be the seventh and final assignment of the semester. As reminder, you are able to drop your two lowest scores.

Now for some code.

